R E P O R T R E S U M E S

ED 020 676 EM 006 298

AN ATTEMPT TO FIND AN A PRIORI MEASURE OF STEP SIZE. COMPARATIVE STUDIES OF PRINCIPLES FOR PROGRAMMING MATHEMATICS IN AUTOMATED INSTRUCTION, TECHNICAL REPORT NO. 13.

BY- ROSEN, ELLEN F. STOLUROW, LAWRENCE M.
ILLINOIS UNIV., URBANA

REPORT NUMBER NDEA-7A-806 PUB DATE JUL 64
EDRS PRICE MF-$0.25 HC-$1.08 25P.

DESCRIPTORS- PROGRAMING PROBLEMS, *PROGRAMED INSTRUCTION, *PROGRAMED TEXTS, LEARNING PROCESSES, LEARNING DIFFICULTIES, *MATHEMATICS INSTRUCTION, MATHEMATICS MATERIALS,

IN ORDER TO FIND A GOOD PREDICTOR OF EMPIRICAL DIFFICULTY, AN OPERATIONAL DEFINITION OF STEP SIZE, TEN PROGRAMER-JUDGES RATED CHANGE IN COMPLEXITY IN TWO VERSIONS OF A MATHEMATICS PROGRAM, AND THESE RATINGS WERE THEN COMPARED WITH MEASURES OF EMPIRICAL DIFFICULTY OBTAINED FROM STUDENT RESPONSE DATA. THE TWO VERSIONS, A 54 FRAME BOOKLET AND A 35 FRAME BOOKLET MADE BY DELETING UNNECESSARY FRAMES FROM THE FIRST BOOKLET, WERE TESTED IN THREE TEACHING MODES. ONE USED THE PROGRAM ONLY, ONE USED CONVENTIONAL TEACHING FOLLOWED BY THE PROGRAM, AND ONE USED THE PROGRAM FOLLOWED BY TEACHING. JUDGES RATED HALF OF EACH PROGRAM, FRAME BY FRAME, ON A TEN POINT SCALE OF FOUR CHARACTERISTICS--THE CONCEPT, THE VEHICLE, THE NUMERAL, AND THE RESPONSE. THE AVERAGE EMPIRICAL DIFFICULTY OF THE TWO VERSIONS DID NOT DIFFER, INDICATING THAT THE STEP SIZE AND NUMBER OF FRAMES DELETED ARE MOST LIKELY NOT IN ONE-TO-ONE CORRESPONDENCE. IT WAS CONCLUDED THAT JUDGES CAN RELIABLY ESTIMATE EMPIRICAL DIFFICULTY AND THAT THE RESPONSE AND NUMERICAL INDICES ARE MOST PROMISING FOR THIS PURPOSE. (BB)

BR TJT6 U11 A
QA.56
#806
A
EM006 298

ED020676

UNIVERSITY OF ILLINOIS
Urbana, Illinois

An Attempt to Find an A Priori Measure of Step Size

Ellen F. Rosen and Lawrence M. Stolurow

COMPARATIVE STUDIES OF PRINCIPLES FOR PROGRAMMING MATHEMATICS IN AUTOMATED INSTRUCTION

Technical Report No. 13
July, 1964

Co-investigators:

Lawrence M. Stolurow
Professor, Department of Psychology
Training Research Laboratory

Max Beberman
Professor, College of Education
University of Illinois Committee
on School Mathematics (UICSM)

Project Sponsor:

Educational Media Branch
U. S. Office of Education
Title VII

Project No. 711151.01

U. S. Office of Education
Title VII

COMPARATIVE STUDIES OF PRINCIPLES FOR PROGRAMMING MATHEMATICS IN AUTOMATED INSTRUCTION

Technical Report No. 13

An Attempt to Find an A Priori Measure of Step Size.

Ellen F. Rosen and Lawrence M. Stolurow

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

July, 1964

EM 006 298

Table of Contents

	Page
List of Tables and Figures	iii
Problem	1
Method	1
Materials	1
Judge's materials	2
Procedure for judges	2
Results	3
Correlations of judgments with empirical difficulty	3
Conclusions	9
Summary	11
Appendix A	12
Appendix B	18
References	20

List of Tables

Table No.		Page
1	Descriptive Statistics on Judges' Ratings of Two Versions of Part 112 of the UICSM Programed Learning materials	4
2	Distribution Statistics for Empirical Difficulty (Student's Response) Under Three Conditions of Use for the Two Versions	5
3	Correlation of Judged and Observed Step Size for the Condition of Use Called Pure Mode (Program Only)	6
4	Correlations of Judged and Observed Step Size for the Condition of Use Called Lead Mode (Program First)	7
5	Correlations of Judged and Observed Step Size for the Condition of the Called Follow Mode (Program Follow)	8

An Attempt to Find An A Priori Measure of Step Size

Ellen F. Rosen and Lawrence M. Stolurow

PROBLEM

Step size in an important determiner of student performance. Although it may seem to be so, step size is not readily measurable. Logically, the most reasonable measure of step size is empirical difficulty as calculated from student performance, but this is an a posteriori measure. An a priori measure is needed. The present investigation is an attempt to find a fine grain predictor of empirical difficulty.

METHOD

Subjects and Judges

The judges who served as raters were ten programers from the staff of UICSM. The subjects (students) have been described elsewhere (Beberman and Stolurow, 1963, Quarterly Report 9 & 10, Chapter VII).

MATERIALS

Student's materials. The materials consisted of the two versions of Part 112[1] of the UICSM-PIP materials (See Beberman and Stolurow, 1963).

[1] Large step size version prepared by Clark Himmel.

Two booklets were prepared for the students' use and were assigned randomly to those available for the study. One version was called the small step version and designated 112S, the other was the large step version and designated 112L.

Both versions were given to students as learning materials under three conditions of use in conjunction with a teacher. In one condition, the program was given to the students, after which the teacher covered the material. This was called the "lead" mode. In a second condition, the program was given to the students, after the teacher had covered the material. This was called the "follow" mode. In the third condition, called the "pure" mode, only the program was given to the student; the teacher did not cover the material.

Judge's materials. Two booklets were prepared for the judges, Judges 1 and Judges 2. These two books consisted of a segment from both student versions so that each judge rated half of each student version.

Procedure for judges. Judges were given one form of the judge's booklets and asked to rate it according to four categories. A copy of the instructions to judges is presented in Appendix A. The instructions are self-explanatory. They define and illustrate the judge's task which was to relate pairs of adjacent steps and to rate changes in complexity on a scale from -5 through +5 on four separate characteristics: (a) the concept; (b) the vehicle; (c) the numeral; and (d) the response.

RESULTS

The judges ratings were converted into standard scores for each category (Guilford, 1956, Pp. 489-494). The standard scores for each step were then averaged across judges within categories and across categories and judges. Thus two sets of ratings were arrived at, one for each (student) booklet version.

From the students' responses an empirical difficulty was calculated (percent of students getting all the problems on the page correct). The means and standard deviations for the ratings and students under the three different conditions of teacher presentation are presented in Table 1 and Table 2, respectively.

Correlations Of Judgments With Empirical Difficulty

Tables 3, 4, and 5 present the correlations of step size judgments with empirical difficulty. The judgments and empirical difficulty were paired by considering the difficulty of the last page of the step as the measure to be predicted. Thus, for example, each judge's ratings of the step from page 1 to page 2 of Part 112, was paired with the empirical difficulty as calculated from students' responses to the questions on page 2 of Part 112. It might be noted here that Part 112 has more than one problem per frame. Consequently these data are likely to have greater reliability than those obtained from more conventional linear programs with only one response per page.

Table 1

Descriptive Statistics on Judges' Ratings of Two Versions of Part 112 of the UICSM Programed Learning Materials

Versions	Category[a]	Mean	Rank	Standard Error	Amount of Change
Part 112S[b]	Concept	-.085	5	.646	
(small step)	Vehicle	.011	1	.593	
	Numeral	.010	2	.654	
	Response	-.004	3	.668	
	Total	-.017	4	.401	
Part 112L[c]	Concept	.172	1	.503	
(large step)	Vehicle	.008	4	.854	
	Numeral	0.000	5	.797	
	Response	.045	3	.696	
	Total	.056	2	.523	

[a]These categories are described in Appendix A.

[b]Based on the average rating of five judges on 51 steps using a standard score conversion of scale values.

[c]Based on the average rating of five judges on 32 steps using a standard score conversion of scale values.

Table 2

Distribution Statistics for Empirical Difficulty (Student's Response) Under Three Conditions of Use for the Two Versions

Version	Conditions of use	Mean Difficulty	Standard Deviation
112S (small step)	Program Lead[a]	78.425	18.705
	Program Follow[b]	75.490	19.007
	"Pure" (Only Program)[c]	75.686	17.740
112L (large step)	Program Lead[d]	78.361	17.983
	Program Follow[e]	76.875	22.141
	"Pure" (Only Program)[f]	74.023	18.229

[a]based on sample of 11 students on 51 pages.

[b]based on sample of 8 students on 51 pages.

[c]based on sample of 20 students on 51 pages.

[d]based on sample of 13 students on 32 pages.

[e]based on sample of 10 students on 32 pages.

[f]based on sample of 16 students on 32 pages.

Table 3

Correlation of Judged and Observed Step Size
for the Condition of Use Called Pure Mode (Program Only)

Version	Concept #	Vehicle #	Numeral #	Response #	Total
Part 112S (51 frames)	-. 080	-. 071	-. 010	-. 278**	-. 178
Part 112L (32 frames)	-. 270	-. 293	-. 329	-. 360*	-. 429*

*for H_o: $\rho = 0$, $r_{.95} = .349$ for 30 df (two-sided).

**for H_o: $\rho = 0$, $r_{.95} \cong .274$ for 49 df (two-sided).

Table 4

Correlations of Judged and Observed Step Size
for the Condition of Use Called Lead Mode (Program First)

Version	Concept	Vehicle	Numeral	Response	Total
Part 112S (51 frames)	-.096	-.127	-.213	-.336**	-.312**
Part 112L (32 frames)	-.248	-.188	-.419*	-.271	-.386*

*for H_o: $\rho = 0$, $r_{.95} = .349$ for 30 df (two-sided).

**for H_o: $\rho = 0$, $r_{.95} \cong .274$ for 49 df (two-sided).

Table 5

Correlations of Judged and Observed Step Size
for the Condition of the Called Follow Mode (Program Follow)

Version					
Part 112 S (51 frames)	-. 031	. 065	. 052	-. 293**	-. 089
Part 112L (32 frames)	-. 089	-. 108	-. 434*	-. 175	-. 289

*for H_o: $\rho = 0$, $r_{.95} = .349$ for 30 df (two-sided).

**for H_o: $\rho = 0$, $r_{.95} \cong .275$ for 49 df (two-sided).

Correlations significantly different from zero at .05 level were obtained from (1) the pure mode (Table 3) between (a) the response category ratings and the empirical difficulty for both the large and small step size programs; and between the overall average (total) rating and difficulty for the large step sequence; (2) the lead mode (Table 4) between (a) the numeral category and difficulty for the large step sequence, (b) the response category and difficulty for the small step sequence, and (c) the average overall rating across categories for both sequences; and (3) the follow mode (Table 5) between the numeral category and difficulty for the large step sequence, and between the response category ratings and difficulty for the small step sequence.

CONCLUSIONS

The results of this study are not exactly clear. A quick glance at Table 2 indicates that, in fact, the average empirical difficulty of the steps did not differ for the two versions within the presentation mode. This is probably due to the fact that the two versions were prepared before the beginning of the study. The large step version was generated by means of deletion of frames which were felt to be unnecessary. Thus, it is quite probably that the two versions really did not differ in terms of step size.

This has potentially important implications for the previous studies of step size (Coulson and Silberman, 1960; Evans, Glaser and Homme, 1960; Glaser and Reynolds, 1962; Maccoby and Sheffield, 1958; Margolius and Sheffield, 1961; Smith and Moore 1961.) in which the typical method of

manipulation has been the simple deletion or addition of frames to create the so-called larger step version. Present results suggest that the deletion procedure may produce an illusion of change other than an actual change in step size. Certainly this simple manipulation is suspect unless step size changes are documented by some additional information relating to program changes produced by frame deletion.

The important point of these results is that step size and number of frames deleted are most likely not in one-to-one correspondence; when aiming at increasing step size one also must consider quality (kind of material deleted) as well as quantity (number or amount of material deleted). This issue of quantity and quality will be discussed in a report on sequential analysis of parts within the sequence and frames within the parts.

The data in Tables 3, 4 and 5 suggest that variations in difficulty probably could be achieved by systematic variation in the response and numeral characteristics of the steps. These two dimensions seem to be the most promising basis for changing step size.

Contrary to the finding of Rothkopf (1963), this study has shown that judges can reliably estimate empirical difficulty by examining the stimulus materials. In part, reliability was obtained, with the present rating scale, by using judgments based upon changes between adjacent frames. The indices that seem to be most promising for this purpose are response and numeral, the former being somewhat more dependable (significant correlations in three out of four possibilities) than the latter (significant correlations in one out of four possibilities).

SUMMARY

This study is an attempt to develop a methodology for the estimation of empirical difficulty under conditions in which the relative range of step sizes is small. The judgment of changes taking place from frame to frame were obtained with a standardized 10 point scale which required the judges to evaluate four characteristics of the stimulus materials: concept, vehicle, numeral and response. Judgments were obtained for a "small-step" version and for the same material with some steps deleted ("large step"). The stimulus materials were booklets consisting of 54 and 35 frames respectively, taken, as a random sample from the original version of the experimental edition of the UICSM High School mathematics programed materials.

APPENDIX A[2]

Instructions for Judges

We are interested in the similarities and differences in pairs of adjacent pages or "learning steps" contained in the accompanying booklet of programed instruction, and we would like your help in finding out how much these adjacent pages are different from and similar to each other with regard to the complexity (abstractness) of certain given characteristics of the material present in the pages. (The pages to be judged will be considered in serial order, i. e., pages 1 and 2 will be compared, then pages 2 and 3, then pages 3 and 4, etc. through the final two pages in booklet.)

We want you to rate the changes in complexity (abstractness) of certain characteristics in going from the first page of the pair to the second page on a scale from -5 through +5, with a rating of zero (0) representing no change in the complexity of a characteristic, ratings above zero representing progressively increasing complexity from the first to the second page, and ratings below zero representing progressively decreasing complexity from the first to the second page, so that a rating of ±5 represents the most expreme change in complexity of a characteristic in either direction. If a characteristic is not present on either of the pages of the pair, record a zero (0) as your rating.

---

[2]Prepared and developed by Clark Himmel to conform to the dimensional requirements developed in work with a program or fractions by L. M. Stolurow with the assistance of Gaila Grubb.

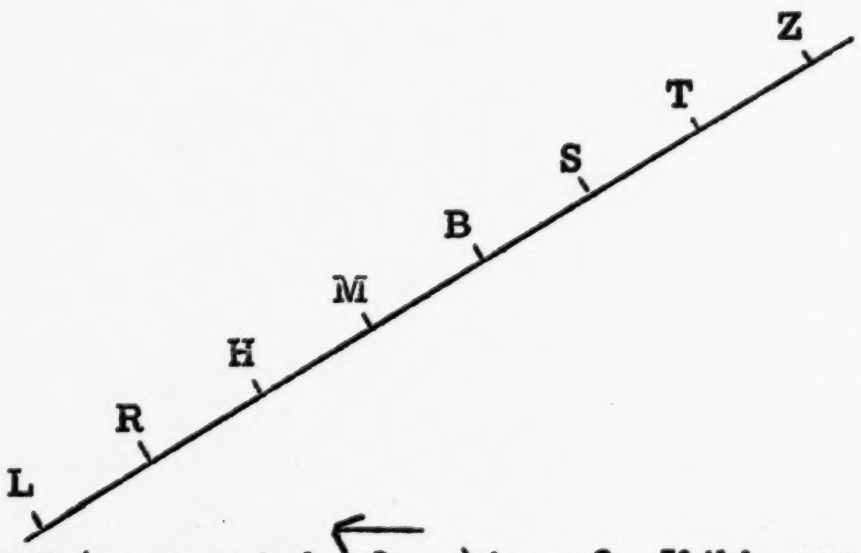
The four characteristics that we want you to consider are (A) the Concept, (B) the Vehicle, (C) the Numeral, and (D) the Response. A description of each of these characteristics, along with an example, and a rating guide is given below.

Concept: refers to the mathematical rule, principle, idea, or closely related group of rules, concepts, conventions, ideas, or principles in mathematics; such as, the associative principle of addition, or the axiomatic system in Euclidean geometry, or the idea of negative numbers.

You should be looking for one of the following: Changes in the complexity, in levels of description or in manner of presentation. You are to identify and rate these changes when leaving one concept and turning to another as they happen within two adjacent pages. Also, note changes in overall complexity when two or more concepts (or, if you prefer, "sub-concepts") are presented simultaneously on one or both of the pair of pages being considered. For example, if only addition is presented on one page and both addition and multiplication are presented on the following page, the change probably is an increase in the complexity of this characteristic. If this occurred then the rating assigned to the pair of pages might be a +2 for the concept.

Vehicle: that which is used to help communicate or convey the concept (and the associated material) being presented by giving a concrete or exemplar background or "real setting" to the problems and expository material; such as, two airplanes traveling toward each other in a rate of travel problem in algebra, or the ledger entries for a retail business in a bookkeeping problem.

This characteristic is one which may not be present on all program steps. Consider the vehicle "a road with mile markers" for presenting the idea of real numbers (both positive and negative), where a trip from R to B (represented $\overrightarrow{3}$) is a +3



and a trip from T to B (represented $\overleftarrow{2}$) is a -2. If this same vehicle with no additions or deletions is present on both pages of a pair, the rating assigned would be zero (0). If it is absent only on the second page of the pair, the rating assigned would be

+5. (The above assumes that no new vehicle characteristics were introduced on either of the pages in the pair.) If something (diagrams, notation, verbal explanation) is added to the vehicle or a new vehicle is introduced in going from the first page to the second, a rating commensurate with the accompanying change in complexity should be assigned. If the same material were deleted from the second page, a rating commensurate with this change should be assigned.

Numeral: refers simply to all symbols for or representations of numbers presented, by the Roman numerals, Hindu-Arabic numerals, or others, plus their accompanying "operators" and "designators," such as +, ÷, $\sqrt{12}$, =, or -7, so that an entire expression like (+16 ÷ -4) x +2 = -8 would be considered under this characteristic.

Consideration should be given to changes in complexity in the types of numerals given on the pages. This should be relatively straightforward, since numerals and their "operators" and "designators" are presented in an explicit notation system. For example, a first page might present addition of simple three digit numerals while the next page calls for multiplication of the square roots of similar three digit numerals. Then the pair would probably receive a fairly high positive rating, perhaps a +3.

Response: refers to the particular answer(s) to be chosen, constructed or written, or in some way indicated by the student as he finishes the problem(s) or question(s) on a page.

Response complexity will vary due to the characteristics of the actual response given and due to the abstractness or difficulty of the specific question(s) or explicitly stated problem(s) to be answered or solved. For example, a response that would be relatively complex in the UICSM Unit I material would be one which is constructed or written by the student; for example, "the associative principle of addition." A relatively less complex response would be choosing one of two alternatives. The second facet of "response" to be considered is the nature of the problem(s) or question(s) to be answered. It also can be scaled in terms of complexity or abstractness. A question like "2 + 2 = ?" is probably less complex than a long and tedious word problem which also requires only a single digit answer.

Each of the characteristics on the pair of steps (pages) to be compared should be rated with regard to the change in complexity (or abstractness in the sense of being abstruse, more difficult to comprehend, ideationally complex or intricate) in going from one step to the next one.

On your rating sheets you will find the four characteristics listed as headings of four columns. Each pair of pages to be compared and then rated is listed at the left. When comparing pairs of pages, do not include the answers and "feedback" material (usually included between the statements "check your answers" and "record your results") in your considerations for rating. We are interested in having you rate the "instructional" and "question" portions of the pages.

Remember:

1. Rate Changes on the scale from

+5	Mid-point 0	-5
Increased Complexity	(no change)	Decreased Complexity

2. Consider the four following characteristics when rating each pair of pages:

   A. Concept

   B. Vehicle

   C. Numeral

   D. Response

3. For each characteristic consider the amount of change in your perception.

APPENDIX B

Sample Rating Sheet

Name

Date

PAGES	Concept	Vehicle	Numeral	Response	
1-2					
2-3					
3-4					
4-5					
5-6					
6-7					
7-8					
8-9					
9-10					
10-11					
11-12					
12-13					
13-14					
14-15					
15-16					
16-17					

Name _____

Date _____

PAGES	Concept	Vehicle	Numeral	Response	
17-18					
18-19					
19-20					
20-21					
21-22					
22-23					
23-24					
24-25					
25-26					
26-27					
27-28					
28-29					
29-30					
30-31					
31-32					
32-33					
33-34					
34-35					
35-36					

Name _____

Date _____

PAGES	Concept	Vehicle	Numeral	Response	
36-37					
37-38					
38-39					
39-40					
41-42					
42-43					
43-44					
44-45					
45					

REFERENCES

Beberman, M. and Stolurow, L. M. Comparative studies of principles for programming mathematics in automated instruction. Semi-Annual Report (Quarterly Reports 9 and 10), USOE, Project No. 711151. 01, June 6, 1963 - December 6, 1963.

Coulson, J. E. and Silberman, H. F. Effects of three variables in a teaching machine. J. educ. Psychol., 1960. 51, 135-43.

Evans, J. L., Glaser, R., and Homme, L. E. An investigation of variation in the properties of verbal learning sequences of a teaching machine type. In A. A. Lumsdaine and R. Glaser (Eds.) Teaching Machines and programed learning: A source book. Washington, D. C.: Dept. of Audiovisual Instruction, Nat. Educ. Assoc., 1960, Pp. 446-51.

Glaser, R. and Reynolds, J. H. Multi-tracking in programmed instruction: Studies with a program on mathematical bases for management decision making. Pittsburgh, Pa.: Univer. of Pittsburgh, Programmed Learning Lab., 1962.

Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1956.

Maccoby, N. and Sheffield, F. D. Theory and experimental research on the teaching of complex sequential procedures by alternate demonstration and practice. In G. Finch and F. Cameron (Eds.) Symposium on Air Force Human Engineering, Personnel and Training Research. Washington, D. C.: Nat. Academy of Science, Nat. Research Council, 1958.

Margolius, G. J. and Sheffield, F. D. Optimum methods of combining practice with filmed demonstration in teaching complex response sequences: serial learning of a mechanical assembly task. In A. A. Lumsdaine (Ed.) Student response in programmed instruction. Washington, D. C.: Nat. Academy of Sciences, Nat. Res. Council, 1961.

Rothlopf, E. Z. Some observations on predicting instructional effectiveness by simple inspection. J. programed Instruction, 1963, 2, (2), 19-20.

Smith, W. I. and Moore, J. W. Size of step and cueing. Psych. Reports, 1962, 10, 287-94.